International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (01), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

SOCIRANK: IDENTIFYING AND RANKING PREVALENT NEWSTOPICS USING SOCIAL MEDIA FACTORS

T.Jeya*[1], V.Madhu Bala[2],

**1 Assistant Professor, Department of Computer science,*
Sri Adi Chunchanagiri Women's College, (India)
2Research Scholar, Department of Computer science,
Sri Adi Chunchanagiri Women's College, (India)

ABSTRACT

To predict interactions between social media and traditional news streams is becoming increasingly relevant for a variety of applications, including: understanding the underlying factors that drive the evolution of data sources, tracking the triggers behind events, and discovering emerging trends. Researchers have developed such interactions by examining volume changes or information diffusions, however, most of them ignore the semantical and topical relationships between news and social media data. Our work is the first attempt to study how news influences social media, or inversely, based on topical knowledge. We introduce a hierarchical Bayesian model that jointly models the news and social media topics and their interactions. We show that our proposed model can capture distinct topics for individual datasets as well as discover the topic influences among multiple datasets. By applying our model to large sets of news and tweets, we demonstrate its significant improvement over baseline methods and explore its power in the discovery of interesting patterns for real world cases.

Keywords:Information filtering, social computing, social network analysis, topic identification, topic ranking.

I.INTRODUCTION

Today, online social media such as Twitter have served as tools for organizing and tracking social events. Understanding the triggers and shifts in opinion driven mass social media data can provide useful insights for various applications in academia, industry, and however, there remains a general lack of finding of what causes the hot spots in social media. Typically, the reasons behind the rapid spread of information can be summarized in terms of two categories: exogenous and endogenous factors. Growing factors are the results of information diffusion inside the social network itself, namely, users obtain information primarily from their online social

network. In contrast, exogenous factors mean that users get information from outside sources first, for example, traditional news media, and then bring it into their social network.

Although previous works have explored both the social media and external news data datasets, few researchers have looked at the endogenous and exogenous factors based on semantical or topical knowledge. They have either sought to identify relevant tweets based on news articles or simply correlated the two data sources through similar patterns in the changing data volume. Still within the same data source, there could be various factors that drive the evolution of information over time. Exogenous factors across multiple datasets make analyzing the evolution and relationship among multiple data streams more difficult. Watching social media and outside news data streams in a united frame can be a practical way of solving this problem. In this paper, we propose a novel topic model, News and Twitter Interaction Topic model (NTIT), that jointly learns social media topics and news topics and subtly capture the influences between topics. The intuition behind this approach is that before a user posts a message, he/she may be influenced either by opinions from his/her online friends or by articles from news agencies. In our new framework, a word in a tweet can be responsive to the topical influences coming either from endogenous factors (tweets) or from exogenous factors (news).

A straightforward approach for identifying topics from different social and news media sources is the application of topic modeling. Many methods have been proposed in this area, such as latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA). Topic modeling is, in essence, the discovery of "topics" in text corpora by clustering together frequently co-occurring words. This approach, however, misses out in the temporal component of prevalent topic detection, that is, it does not take into account how topics change with time. Furthermore, topic modeling and other topic detection techniques do not rank topics according to their popularity by taking into account their prevalence in both news media and social media.

We introduce an unsupervised system—SociRank—which effectively identifies news topics that are prevalent in both social media and the news media, and then ranks them by relevance using their degrees of MF, UA, and UI. Even though this paper focuses on news topics, it can be easily adapted to a wide variety of fields, from science and technology to culture and sports. To the best of our knowledge, no other work attempts to employ the use of either the social media interests of users or their social relationships to aid in the ranking of topics. Moreover, SociRank undergoes an empirical framework, comprising and integrating several techniques, such as keyword extraction, measures of similarity, graph clustering, and social network analysis. The effectiveness of our system is validated by extensive controlled and uncontrolled experiments.

II.EMERGENCE OF TWITTER AS A NEWS MEDIA

Computer science research community has analyzed relevance of online social media, in particular Twitter, as news disseminating agent. Kwak et al. showed the prominence of Twitter as a news media, they showed that 85% topics discussed on Twitter are related to news. Their work highlighted the relationship between user specific parameters v/s the tweeting activity patterns, like analysis of the number of followers and followees v/s the tweeting (re-tweeting) numbers. Zhao et al. in their work, used unsupervised topic modeling to compare the news topic from Twitter versus New York Times (a traditional news dissemination medium). They showed that

Twitter users are relatively less interested in world news, still they are active in spreading news of important world events. Lu et al. showed how tweets related to news event on Twitter can be mapped using energy function. The methods proposed act like novel event detection techniques. The study analyzed 900 news events through 2010-2011. Castillo et al. performed qualitative and quantitative analysis on online social media activity about news articles. They concluded that news articles describing breaking news events have more repetitive social media reactions, than in-depth articles.

III.ANALYZING TWITTER DATA DURING REAL-WORLD EVENTS

The posts and activity on Twitter, impacts and plays a vital role in various real world events. Role of Twitter has been analyzed by computer scientists, psychologists and sociologists for impact in the real-world. Twitter has progressed from being merely a medium to share users' opinions; to an information sharing and dissemination agent; to propagation and coordination of relief and response efforts. Some of the popular case studies analyzed by computer scientists have been, Twitter activities during elections, natural disasters (like hurricanes, wildfires, floods, etc.), political and social uprisings (like Libya and Egypt crisis) and terrorist attacks (like Mumbai triple bomb blasts). Content and user activity patterns of Twitter during events have been analyzed for both positive and negative aspects. Some of the problems studied that result in bad quality of data, presence of spam and phishing posts, content spreading rumors / fake news, privacy breach of users via the content shared by them and use of Twitter for propagation and instigation of hate among people. Researchers have used machine learning, information retrieval, social network analysis and image and video analysis for the purpose of analyzing and characterizing Twitter usage during real-world events.

We introduce some of the research work done in applying user modeling techniques to analyze behavior of users on social networks. Yin et al. modeled user behavior using two factors: the topics related to users' intrinsic interests and the topics related to temporal context . They created a latent class statistical mixture model, called Dynamic Temporal Context-Aware Mixture model (DTCAM). They evaluated their system on four large-scale social media datasets. The authors demonstrated how user modeling techniques can be effectively used to improve the performance of recommender systems for social networks. Xu et al. introduced a mixed latent topic model to combine various factors to model users' posting behavior on Twitter. The authors assumed that a user's behavior is influenced by three factors: breaking news, posts from social friends and user's interest. They developed and showed that their model outperforms other user models in handling the perplexity of held-out content and the quality of generated latent topics. Abel et al. developed a user modeling framework for news recommendations on Twitter using more than 2 million tweets. The authors proposed different strategies for creating hash tag-based, entity based or topic-based user profiles using semantic enrichment and temporal factors. Their results showed that consideration of temporal profile patterns can improve recommendation quality.

IV.LITERATURE REVIEW

Much research has been carried out in the field of topic identification—referred to more formally as topic modeling. Two traditional methods for detecting topics are LDA [1] and PLSA [2], [3]. LDA is a generative probabilistic model that can be applied to different tasks, including topic identification. PLSA, similarly, is a statistical technique, which can also be applied to topic modeling. In these approaches, however, temporal information is lost, which is paramount in identifying prevalent topics and is an important characteristic of social media data. Furthermore, LDA and PLSA only discover topics from text corpora; they do not rank based on popularity orprevalence. Wartena and Brussee [4] implemented a method to detect topics by clustering keywords. Their method entails the clustering of keywords—based on different similarity measures—using the inducedk-bisecting clustering algorithm [5]. Although they do not employ the use of graphs, they do observe that a distance measure based on the Jensen–Shannon divergence (or information radius [6]) of probability distributions performs well. More recently, research has been conducted in identifying topics and events from social media data, taking into account temporal information. Cataldiet al.[7] proposed a topic detection technique that retrieves real-time emerging topics from Twitter. Their method uses the set of terms from tweets and model their life cycle according to a novel aging theory. Additionally, they take into account social relationships—more specifically, the authority of the users in the network—to determine the importance of the topics. Zhaoet al.[8] carried out similar work by developing a Twitter-LDA model designed to identify topics in tweets. Their work, however, only considers the personal interests of users, and not prevalent topics at a global scale. Another trending area of related research is the detection of "bursty" topics (i.e., topics or events that occur in short, sudden episodes). Diao et al. [9] proposed a method that uses a state machine to detect bursty topics in microblogs. Their method also determines whether user posts are personal or refer to a particular trending topic. Yin et al.[10] also developed a model that detects topics from social media data, distinguishing between temporal and stable topics. These methods, however, only use data from microblogs and do not attempt to integrate them with real news. Additionally, the detected topics are not ranked by popularity or prevalence.

Wanget al.[11] proposed a method that takes into account the users' interest in a topic by estimating the amount of times they read stories related to that particular topic. They refer to this factor as the UA. They also used an aging theory developed by Chenet al.[12] to create, grow, and destroy a topic. The life cycles of the topics are tracked by using an energy function. The energy of a topic increases when it becomes popular and it diminishes over time unless it remains popular. We employ variants of the concepts of MF and UA to meet our needs, as these concepts are both logical and effective. Other works have made use of Twitter to discover news-related content that might be considered important. Sankaranarayanan et al. [13] developed a system called TwitterStand, which identifies tweets that correspond to breaking news. They accomplish this by utilizing a clustering approach for tweet mining. Phelan et al. [14] developeda recommendation system that generates a ranked list of news stories. News are ranked based on the co-occurrence of popular terms within the users' RSS and Twitter feeds. Both of these systems aim to identify emerging topics, but give no insight into their popularity over time. Moreover, the work by Phelan et al. [14] only produces a personalized ranking (i.e., news

articles tailored specifically to the content of a single user), rather than providing an overall ranking based on a sample of all users. Nevertheless, these works provide us with a basis for extending the premise of UA. Research has also been carried out in topic discovery and ranking from other domains. Shubhankar et al. [15] developed an algorithm that detects and ranks topics in a corpusof research papers. They used closed frequent keyword-sets to form topics and a modification of the Page Rank [16] algorithm to rank them. Their work, however, does not integrate or collaborate with other data sources, as accomplished by SociRank.

V.CONCLUSION

Our model includes jointly topic modeling on multiple data sources in an asymmetrical frame, which benefits the modeling performance for both long and short texts. We present the results of applying model to two large-scale datasets and show its effectiveness over non-trivial baselines. Based on the outputs of model, further efforts are made to understand the complex interaction between news and social media data. Through extensive experiments, we find following factors: 1) even for the same events, focuses of news and Twitter topics could be greatly different; 2) topic usually occurs first in its dominant data source, but occasionally topic first appearing in one data source could be a dominant topic in another dataset; 3) generally, news topics are much more influential than Twitter topics.

REFERENCES

[1] *D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022,* Jan. 2003.

[2] T. Hofmann, "*Probabilistic latent semantic analysis*," inProc. 15th Conf. Uncertainty Artif. Intell., 1999, pp. 289–296.

[3] T. Hofmann, "*Probabilistic latent semantic indexing*," in Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, Berkeley, CA, USA, 1999, pp. 50–57.

[4] C. Wartena and R. Brussee, "*Topic detection by clustering keywords*," inProc. 19th Int. Workshop Database Expert Syst. Appl. (DEXA),Turin, Italy, 2008, pp. 54–58.

[5] F. Archetti, P. Campanelli, E. Fersini, and E. Messina, "*A hierarchical document clustering environment based on the induced bisecting k-means*," in Proc. 7th Int. Conf. Flexible Query Answering Syst., Milan, Italy, 2006, pp. 257–269.

[6] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[7] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging *topic detection on Twitter based on temporal and social terms evaluation*," in Proc. 10th Int. Workshop Multimedia Data Min. (MDMKDD),Washington, DC, USA, 2010.

[8] W. X. Zhaoet al., "*Comparing Twitter and traditional media using topic models*," inAdvances in Information Retrieval. Heidelberg, Germany: Springer Berlin Heidelberg, 2011, pp. 338–349.

[9] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, "*Finding bursty topics from microblogs*," inProc. 50th Annu. Meeting Assoc. Comput. Linguist. Long Papers, vol. 1. 2012, pp. 536–544.

[10] H. Yin, B. Cui, H. Lu, Y. Huang, and J. Yao, "*A unified model for stable and temporal topic detection from social media data*," in Proc. IEEE 29th Int. Conf. Data Eng. (ICDE), Brisbane, QLD, Australia, 2013,pp. 661–672.

[11] C. Wang, M. Zhang, L. Ru, and S. Ma, "*Automatic online news topic ranking using media focus and user attention based on aging theory*," in Proc. 17th Conf. Inf. Knowl. Manag., Napa County, CA, USA, 2008,pp. 1033–1042.

[12] C. C. Chen, Y.-T. Chen, Y. Sun, and M. C. Chen, "*Life cycle modeling of news events using aging theory,"* inMachine Learning: ECML 2003. Heidelberg, Germany: Springer Berlin Heidelberg, 2003,pp. 47–59.

[13] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "*TwitterStand: News in tweets,"* in Proc. 17th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst., Seattle, WA, USA,2009, pp. 42–51.

[14] O. Phelan, K. McCarthy, and B. Smyth, "*Using Twitter to recommend real-time topical news*," inProc. 3rd Conf. Recommender Syst., New York, NY, USA, 2009, pp. 385–388.[15] K. Shubhankar, A. P. Singh, and V. Pudi, "*An efficient algorithm for topic ranking and modeling topic evolution*," in Database Expert Syst.Appl., Toulouse, France, 2011, pp. 320–330.

[16] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine,"Comput. Network., vol. 56, no. 18, pp. 3825–3833,2012.